تلاش‌هایی برای پیدا کردن کران بالایی بهتر برای مسئله بندیت پیوسته و لیپ‌شیتز

درس: آنالیز احتمالاتی ابعاد بالا

تهیه‌کنندگان:
مهدی دارابی
حامی عبادزاده سمنانی

استاد: دکتر محمدحسین یاسائی میبدی
دانشگاه صنعتی شریف
۴ اسفند ۱۴۰۴

فهرست مطالب

۱ مقدمه ۲

۲ تعریف مسئله (حالت گسسته و پاداش‌های کران‌دار) ۲

۳ نمونه‌ای از سیاست‌های یادگیری در بندیت چند بازو ۲

۱.۳ الگوریتم UCB۱ . . . . . ۲

۲.۳ کران پشیمانی سیاست UCB۱ . . . . . ۳

۴ مسئله پیشنهادی ۳

۱.۴ مسئله‌ی بندیت پیوسته با پاداشِ کران‌دار . ۳

۲.۴ فرض لیپ‌شیتز بودنِ f و توجیه آن . . . ۳

۳.۴ گسسته‌سازی فضا با ϵـ پوشش ۴

۴.۴ مسئله پیشنهادی ۴

۵ تلاش‌ها ۴

۱.۵ تلاش اول : گسسته‌سازی با ϵـ پوشش و کران مستقل از فاصله ۴

۱.۱.۵ سیاست «گسسته‌سازی + UCB1» ۴

۲.۱.۵ گام اول: کنترل خطای گسسته‌سازی با لیپ‌شیتز ۴

۳.۱.۵ گام دوم: کران مستقل از فاصله برای UCB1 روی K بازو . . . ۵

۴.۱.۵ ترکیب دو گام و بهینه‌سازی ϵ . . ۵

۲.۵ تلاش دوم (ناموفق): بازنویسی پشیمانی به‌صورت کران از بالا برای امیدِ سوپریمم یک فرایند ۵

۱.۲.۵ ایده‌ی کلی و مانع اصلی ۵

۲.۲.۵ یک مثال «بد» برای شهود . . . ۵

۳.۲.۵ کران بدیهیِ پشیمانی ۶

۴.۲.۵ جمع‌بندی نتیجه‌ی این تلاش . . ۶

پیوست‌ها ۶

به نام خدا

۱ مقدمه

در بسیاری از مسائل تصمیم‌گیری، باید به‌صورت تکرارشونده یکی از چند گزینه را انتخاب کنیم، در حالی‌که کیفیت واقعی هر گزینه از پیش معلوم نیست و فقط با آزمون و مشاهده پیامدها آشکار می‌شود. مسئله‌ی Multi-Armed Bandit یک چارچوب ساده برای مدل‌سازی همین وضعیت است و هسته‌ی اصلی آن «موازنه‌ی اکتشاف/بهره‌برداری» است: از یک سو باید گزینه‌های کمترشناخته را امتحان کرد تا اطلاعات به‌دست آید (اکتشاف)، و از سوی دیگر باید تا حد ممکن از گزینه‌ای استفاده کرد که تاکنون بهتر عمل کرده است (بهره‌برداری).

این چارچوب در کاربردهای متعددی ظاهر می‌شود؛ برای مثال در تبلیغات و A/B-test، سامانه‌های پیشنهاددهنده، بهینه‌سازی برخط تجربه کاربری، و تخصیص منابع در شرایط عدم‌قطعیت. هدف، به‌طور کلی، بیشینه‌سازی عملکرد تجمعی در طول زمان است، در حالی‌که یادگیری نیز هم‌زمان با تصمیم‌گیری رخ می‌دهد.

۲ تعریف مسئله (حالت گسسته و پاداش‌های کران‌دار)

تعریف این بخش مطابق مقاله‌ی Auer, Cesa-Bianchi, and Fischer (2002) است [۱] و ممکن است در برخی منابع دیگر متفاوت ارائه شود.

مسئله‌ی Multi-Armed Bandit با K بازو به‌صورت زیر مدل می‌شود. برای هر بازو $i \in \{1, \dots, K\}$، دنباله‌ی پاداش‌ها

$$Y_{i,1}, Y_{i,2}, \dots$$

مستقل و هم‌توزیع با توزیع ناشناخته‌ی P_i هستند و

$$\mu_i \;:=\; \mathbb{E}[Y_{i,1}]$$

امید پاداش بازوی i را نشان می‌دهد. همچنین دنباله‌های پاداشِ بازوهای مختلف نسبت به هم مستقل فرض می‌شوند. در این مقاله پاداش‌ها کران‌دارند؛ یعنی برای هر i،

$$\operatorname{supp}(P_i) \subseteq [0, 1].$$

یک سیاست یا Policy در هر گام $t = 1, 2, \dots$ یک بازو را انتخاب می‌کند. اگر $T_i(n)$ تعداد دفعات انتخاب بازوی i تا پایان گام n باشد، آنگاه پاداش تجمعیِ مورد انتظار برابر است با

$$\mathbb{E}\left[\sum_{i=1}^{K} \sum_{s=1}^{T_i(n)} Y_{i,s}\right] \;=\; \sum_{i=1}^{K} \mu_i \, \mathbb{E}[T_i(n)].$$

بازوی بهینه را با

$$\mu^* \;:=\; \max_{1 \le i \le K} \mu_i$$

تعریف می‌کنیم. و در نهایت پشیمانی یا Regret پس از n گام برابر است با

$$R(n) \;:=\; \mu^* n \;-\; \sum_{i=1}^{K} \mu_i \, \mathbb{E}[T_i(n)].$$

هدف مسئله: طراحی سیاستی که regret را کمینه کند؛ یعنی برای افق n تا حد امکان مقدار $R(n)$ کوچک باشد.

۳ نمونه‌ای از سیاست‌های یادگیری در بندیت چند بازو

۱.۳ الگوریتم UCB۱

الگوریتم UCB1 که در [۱] معرفی شده است، برای هر بازو در هر گام یک «شاخص» تعریف می‌کند و بازویی را انتخاب می‌کند که این شاخص بیشترین مقدار را داشته باشد. اگر تا گام $t-1$، بازوی i تعداد $T_i(t-1)$ بار انتخاب شده باشد، میانگین تجربی آن برابر است با

$$\widehat{\mu}_i(t-1) = \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} Y_{i,s}.$$

در گام t شاخص بازوی i به‌صورت

$$I_i(t) = \widehat{\mu}_i(t-1) + \sqrt{\frac{2 \ln t}{T_i(t-1)}}$$

تعریف می‌شود و بازویی انتخاب می‌شود که $I_i(t)$ را بیشینه کند.

جمله‌ی اول بیانگر برآورد فعلی کیفیت بازو (بهره‌برداری) و جمله‌ی دوم بیانگر میزان عدم‌قطعیت نسبت به آن (اکتشاف) است. هرچه بازویی کمتر آزموده شده باشد، جمله‌ی دوم بزرگ‌تر است و الگوریتم آن را برای جمع‌آوری اطلاعات بیشتر ترجیح می‌دهد. در آغاز اجرا، هر بازو یک‌بار انتخاب می‌شود تا میانگین تجربی آن تعریف گردد.

۲.۳ کران پشیمانی سیاست UCB۱

قضیه ۱.۳. *فرض کنید پاداش‌ها در بازه‌ی $[0,1]$ کران‌دار باشند. اگر از الگوریتم UCB1 استفاده شود، آنگاه برای هر بازوی زیربهینه i یعنی $\Delta_i := \mu^* - \mu_i > 0$، تعداد انتخاب‌های مورد انتظار آن تا گام n دارای کران زیر است:*

$$\mathbb{E}[T_i(n)] \le \frac{8\ln n}{\Delta_i^2} + 1 + \frac{\pi^2}{3}.$$

در نتیجه، پشیمانی کل دارای کران لگاریتمی زیر خواهد بود:

$$R(n) \le \sum_{i:\Delta_i>0} \left(\frac{8\ln n}{\Delta_i} + \left(1+\frac{\pi^2}{3}\right)\Delta_i \right).$$

۴ مسئله پیشنهادی

قبل از بیان مسئله پیشنهادی نیاز به تعریف چند مطلب داریم.

۱.۴ مسئله‌ی بندیت پیوسته با پاداشِ کران‌دار

در بندیت پیوسته، فضای عمل یک مجموعه‌ی $\mathcal{X}$ است (عموما $\mathcal{X} \subset \mathbb{R}^d$). در هر گام $t = 1, 2, \ldots$ عامل یک عمل انتخاب می‌کند

$$X_t \in \mathcal{X},$$

و سپس پاداش تصادفیِ کران‌دار مشاهده می‌شود

$$Y_t \in [0,1].$$

تابع میانگین پاداش به‌صورت

$$f : \mathcal{X} \to [0,1], \qquad f(x) := \mathbb{E}[\,Y_t \mid X_t = x\,]$$

تعریف می‌شود. همچنین

$$f^* := \sup_{x\in\mathcal{X}} f(x)$$

و پشیمانی مورد انتظار پس از n گام برابر است با

$$R(n) := \sum_{t=1}^{n} \mathbb{E}[f^* - f(X_t)]\,.$$

هدف، مانند قبل طراحی سیاستی است که $R(n)$ را کمینه کند.

۲.۴ فرض لیپ‌شیتز بودنِ f و توجیه آن

در بسیاری از کارهای مربوط به بندیت پیوسته، علاوه بر کران‌دار بودن پاداش، فرض می‌شود تابع میانگین پاداش f روی $\mathcal{X}$ لیپ‌شیتز است؛ یعنی عددی $L > 0$ وجود دارد به‌طوری‌که برای هر $x, x' \in \mathcal{X}$،

$$|f(x) - f(x')| \le L \times d(x, x').$$

این فرض بیان می‌کند که «اعمال نزدیک» باید میانگین پاداش نزدیک داشته باشند.

با وجود اینکه این توابع **بسیار بسیار** محدود هستند، این فرض در برخی کاربردها نامعقول **نیست**. فرض کنید کنش یک «قیمت» باشد:

$$x \in [0,1].$$

در هر مرحله، شما قیمت x را اعلام می‌کنید و «درآمد» تصادفی زیر را مشاهده می‌کنید:

$$Y \;=\; x \cdot \mathbf{1}\{\text{مشتری با قیمت x خرید می کند}\}.$$

تابع پاداشِ میانگین (میانگین شرطی پاداش) را به صورت زیر تعریف می‌کنیم:

$$f(x) \;=\; \mathbb{E}[Y \mid X = x] \;=\; x\,p(x),$$

که در آن $p(x)$ احتمال خرید در قیمت x است.

در بسیاری از بازارها، تغییرات کوچک در قیمت معمولاً باعث «جهش‌های ناگهانی» در احتمال خرید نمی‌شود؛ به بیان دیگر، $p(x)$ (و در نتیجه $f(x)$) معمولاً با تغییرات کوچک x به‌طور تدریجی تغییر می‌کند. این شهود همان چیزی است که فرضِ لیپ‌شیتز برای f بیان می‌کند.

۳.۴ گسسته‌سازی فضا با ϵـ پوشش

یک راه برای استفاده از یک سیاست گسسته در مسئله‌ی بندیت پیوسته این است که یک ϵـ پوشش از $\mathcal{X}$ بسازیم. فرض کنید $(\mathcal{X}, d)$ یک فضای متریک است و $N_\epsilon \subseteq \mathcal{X}$ یک ϵـ پوشش متناهی باشد؛ یعنی برای هر $x \in \mathcal{X}$ نقطه‌ای $\tilde{x} \in N_\epsilon$ وجود دارد به‌طوری‌که $d(x, \tilde{x}) \leq \epsilon$. در این صورت با تعریف

$$K := |N_\epsilon|, \qquad N_\epsilon = \{x_1, \ldots, x_K\},$$

می‌توان نقاط $x_1, \ldots, x_K$ را به‌عنوان بازوهای یک مسئله‌ی گسسته در نظر گرفت و برای نمونه الگوریتم UCB1 را روی این مجموعه اجرا کرد.

۴.۴ مسئله پیشنهادی

پرسش اصلی. اگر در مسئله‌ی بندیت پیوسته با پاداش کران‌دار، تابع میانگین پاداش f لیپ‌شیتز باشد، آیا می‌توان با *گسسته‌سازیِ فضای عمل* از طریق یک ϵـ پوشش و اجرای مستقیم الگوریتم UCB1 روی مجموعه‌ی N_ϵ، به کران پشیمانی‌ای دست یافت که (پس از انتخاب مناسب ϵ) از کران ارائه‌شده در قضیه‌ی مربوط به حالت پیوسته (قضیه‌ی ۱) بهتر باشد؟

۵ تلاش‌ها

۱.۵ تلاش اول : گسسته‌سازی با ϵـ پوشش و کران مستقل از فاصله

در این بخش «تلاش اول» پروژه را به‌صورت دقیق صورت‌بندی و کران بالا را استخراج می‌کنیم. فرض می‌کنیم $\mathcal{X} = [0,1]^d$ و پاداش‌ها کران‌دارند ($Y_t \in [0,1]$). همچنین تابع میانگین پاداش $f : \mathcal{X} \to [0,1]$ نسبت به متریک $d(\cdot,\cdot)$ لیپ‌شیتز با ثابت L است:

$$|f(x) - f(x')| \leq L\, d(x, x') \qquad (\forall x, x' \in \mathcal{X}).$$

(برای آنکه اندازه‌ی پوشش دقیقاً از مرتبه‌ی ϵ^{-d} شود، می‌توان d را به‌عنوان $d(x,x') = \|x - x'\|_\infty$ در نظر گرفت.)

۱.۱.۵ سیاست «گسسته‌سازی + UCB1»

یک ϵـ پوشش متناهی $N_\epsilon = \{x_1, \ldots, x_K\} \subset \mathcal{X}$ انتخاب می‌کنیم، یعنی برای هر $x \in \mathcal{X}$ نقطه‌ای $\tilde{x} \in N_\epsilon$ وجود دارد که $d(x, \tilde{x}) \leq \epsilon$. سپس در هر گام t تنها یکی از نقاط شبکه را انتخاب می‌کنیم:

$$X_t \in N_\epsilon,$$

و الگوریتم UCB1 را روی K بازوی متناظر با نقاط $x_1, \ldots, x_K$ اجرا می‌کنیم. میانگین بازوی i برابر خواهد بود با:

$$\mu_i = f(x_i).$$

پشیمانی نسبت به بهینه‌ی پیوسته را به‌صورت

$$R(n) := \sum_{t=1}^{n} \mathbb{E}[f^* - f(X_t)]\,,$$
$$f^* := \sup_{x \in \mathcal{X}} f(x).$$

تعریف می‌کنیم.

۲.۱.۵ گام اول: کنترل خطای گسسته‌سازی با لیپ‌شیتز

لم ۱ (خطای تقریب شبکه). *اگر N_ϵ یک ϵـ پوشش باشد، و f با ثابت L لیپ‌شیتز باشد، آنگاه*

$$0 \leq f^* - f^*_\epsilon \leq L\epsilon, \qquad f^*_\epsilon := \max_{x \in N_\epsilon} f(x).$$

اکنون پشیمانی را به دو جزء می‌شکنیم:

$$R(n) = \sum_{t=1}^{n} (\mathbb{E}[f^* - f^*_\epsilon]) \;+\; \sum_{t=1}^{n} (\mathbb{E}[f^*_\epsilon - f(X_t)])$$
$$\leq n \times (\mathbb{E}[f^* - f^*_\epsilon]) + R^\epsilon_n,$$

که در آن

$$R^\epsilon_n := \sum_{t=1}^{n} (\mathbb{E}[f^*_\epsilon - f(X_t)])$$

پشیمانی نسبت به بهترین نقطه روی شبکه است. بنابراین با لم ۱ داریم:

$$R(n) \leq L\epsilon n + R^\epsilon_n.$$

۳.۱.۵ گام دوم: کران مستقل از فاصله برای UCB1 روی K بازو

در مسئله‌ی گسسته با K بازو و پاداش‌های $[0,1]$، الگوریتم UCB1 مطابق [۱] برای هر بازوی زیربهینه i با

$$\Delta_i := \mu^* - \mu_i > 0$$

تضمین می‌کند:

$$\mathbb{E}[T_i(T)] \le \frac{8\ln T}{\Delta_i^2} + 1 + \frac{\pi^2}{3}. \qquad (\star)$$

از این رابطه می‌توان یک کران «مستقل از فاصله» استخراج کرد.

لم ۲ (کران مستقل از فاصله برای UCB1). *در مسئله‌ی گسسته با K بازو و پاداش‌های کران‌دار در $[0,1]$، پشیمانی مورد انتظار UCB1 پس از n گام ارضا می‌کند:*

$$R_n^\epsilon \le 4\sqrt{2K_\epsilon n\ln(n)} + \Big(1+\frac{\pi^2}{3}\Big)K_\epsilon.$$

۴.۱.۵ ترکیب دو گام و بهینه‌سازی ϵ

برای $\mathcal{X} = [0,1]^d$ (با متریک $\|\cdot\|_\infty$)، یک ϵـ پوشش متناهی با

$$K_\epsilon = |N_\epsilon| \le \Big\lceil \frac{1}{\epsilon}\Big\rceil^d \le \Big(\frac{1}{\epsilon}+1\Big)^d$$

وجود دارد؛ در نتیجه از مرتبه‌ی $K \lesssim \epsilon^{-d}$ است. با ترکیب لم ۱ و لم ۲ به دست می‌آوریم:

$$R(n) \le L\epsilon n + 4\sqrt{2K_\epsilon n\ln(n)} + \Big(1+\frac{\pi^2}{3}\Big)K_\epsilon.$$

با جایگذاری $K_\epsilon \approx \epsilon^{-d}$ و انتخاب

$$\epsilon := n^{-\frac{1}{d+2}}, \qquad \Rightarrow \qquad K_\epsilon = \epsilon^{-d} = n^{\frac{d}{d+2}},$$

داریم:

$$L\epsilon n = Ln^{\frac{d+1}{d+2}},$$
$$4\sqrt{2K_\epsilon n\ln(n)} = 4\sqrt{2\ln(n)}\, n^{\frac{d+1}{d+2}},$$
$$\Big(1+\frac{\pi^2}{3}\Big)K_\epsilon = \Big(1+\frac{\pi^2}{3}\Big)n^{\frac{d}{d+2}}.$$

پس در نهایت:

$$R(n) \le L\, n^{\frac{d+1}{d+2}} + 4\sqrt{2\ln(n)}\, n^{\frac{d+1}{d+2}} + \Big(1+\frac{\pi^2}{3}\Big)n^{\frac{d}{d+2}}.$$

این همان کرانی است که در «تلاش اول» به آن رسیدیم.

۲.۵ تلاش دوم (ناموفق): بازنویسی پشیمانی به‌صورت کران از بالا برای امیدِ سوپریمم یک فرایند

انگیزه‌ی این تلاش استفاده از لمِ زیر بود که برای یک فرایند تصادفی لیپ‌شیتز و زیرگاوسی، امیدِ سوپریمم را با عدد پوشش کنترل می‌کند. به طور خلاصه، اگر $\{Z_x\}_{x\in\mathcal{X}}$ فرایندی لیپ‌شیتز و زیرگاوسی باشد، آنگاه کرانی از شکل

$$\mathbb{E}\Big[\sup_{x\in\mathcal{X}} Z_x\Big] \le \varepsilon\,\mathbb{E}[C] + \sqrt{2\sigma^2\log N(\mathcal{X},d,\varepsilon)}$$

برقرار است که در آن C ضریب لیپ‌شیتز و σ پارامتر زیرگوسی فرآیند است.

۱.۲.۵ ایده‌ی کلی و مانع اصلی

برای به‌کارگیری چنین لمی، باید پشیمانیِ افق ثابت n را به شکلی از نوع

$$R(n) \le \mathbb{E}\Big[\sup_{x\in\mathcal{X}} Z_x\Big] \qquad \text{یا حداقل} \qquad \mathbb{E}\Big[\sup_{x\in\mathcal{X}} Z_x\Big]$$

کاهش داد؛ به‌گونه‌ای که $\{Z_x\}_{x\in\mathcal{X}}$ (۱) نسبت به x لیپ‌شیتز باشد و (۲) زیرگاوسی باشد.

۲.۲.۵ یک مثال «بد» برای شهود

برای ایجاد شهود، می‌توان به‌طور صوری کمیتی مانند

$$Z_x := nf(x) - \sum_{t=1}^n f(X_t)$$

را در نظر گرفت. این کمیت از این جهت وسوسه‌انگیز است که اگر $x^* \in \arg\max_{x\in\mathcal{X}} f(x)$ باشد، آنگاه

$$Z_{x^*} = nf(x^*) - \sum_{t=1}^n f(X_t)$$

همان «پشیمانی شبه‌پاداش» (با جایگزینی Y_t به جای $f(X_t)$) را بازتولید می‌کند. همچنین اگر f لیپ‌شیتز باشد، برای هر $x, x' \in \mathcal{X}$ داریم

$$|Z_x - Z_{x'}| = n\,|f(x) - f(x')| \le (n\,L)\,d(x, x'),$$

پس $\{Z_x\}$ نسبت به متریک d لیپ‌شیتز است.

همچنین توجه کنید، Z_x کران‌دار هم هست چرا که $f(x) \in [0,1]$ و هر جمله‌ی $f(X_t)$ نیز در همین بازه است، برای هر x داریم

$$-n \;\le\; Z_x = n\,f(x) - \sum_{t=1}^{n} f(X_t) \;\le\; n,$$

بنابراین $\{Z_x\}_{x\in\mathcal{X}}$ یک فرایند کران‌دار (و در نتیجه زیرگاوسی) است.

با این حال، این انتخاب مشکلی اساسی دارد چرا که Z_x مرکزی نیست، یعنی به طور کلی

$$\mathbb{E}[Z_x] = n\,f(x) - \sum_{t=1}^{n} \mathbb{E}\big[f(X_t)\big] \neq 0,$$

و این با یکی از شرایط لازم در لمِ یادشده (که روی فرایندهای زیرگاوسیِ مرکزدار اعمال می‌شود) سازگار نیست.

اگر بخواهیم آن را مرکز کنیم، می‌توانیم

$$\tilde{Z}_x := Z_x - \mathbb{E}[Z_x]$$

را تعریف کنیم. در این صورت

$$\begin{aligned}\tilde{Z}_x &= \big(n\,f(x) - \sum_{t=1}^{n} f(X_t)\big) \\ &\quad - \big(n\,f(x) - \sum_{t=1}^{n} \mathbb{E}[f(X_t)]\big) \\ &= -\sum_{t=1}^{n} f(X_t) + \sum_{t=1}^{b} \mathbb{E}[f(X_t)],\end{aligned}$$

که دیگر به x وابسته نیست. بنابراین

$$\sup_{x\in\mathcal{X}} \tilde{Z}_x = \tilde{Z}_x \quad \text{و} \quad \mathbb{E}\Big[\sup_{x\in\mathcal{X}} \tilde{Z}_x\Big] = \mathbb{E}[\tilde{Z}_x] = 0.$$

در نتیجه، به‌کارگیری آن لم در بهترین حالت فقط یک کرانِ صفر برای کمیتِ مرکزی $\mathbb{E}[\sup_x \tilde{Z}_x]$ می‌دهد، در حالی‌که پشیمانیِ مورد نظر ما به $\sup_x Z_x$ مربوط است و این رویکرد عملاً کران مفیدی برای پشیمانی فراهم نمی‌کند.

۳.۲.۵ کران بدیهیِ پشیمانی

از آنجا که پاداش‌ها کران‌دارند، پشیمانی در هر گام حداکثر 1 است؛ بنابراین

$$0 \le \sum_{t=1}^{n} \big(f^* - f(X_t)\big) \le n, \qquad \Rightarrow \qquad R(n) \le n.$$

این بدترین کران عمومیِ ممکن (بدون استفاده از ساختار بیشتر) است.

۴.۲.۵ جمع‌بندی نتیجه‌ی این تلاش

در نهایت، با وجود تلاش برای ساختن یک فرایند $\{Z_x\}_{x\in\mathcal{X}}$ که هم لیپ‌شیتز و هم زیرگاوسی باشد و بتوان پشیمانی را به امیدِ سوپریمم آن مرتبط کرد، کران‌های حاصل از کاربرد لمِ فوق در بهترین حالت به نرخ مرتبه‌ی $\mathcal{O}(n)$ منجر می‌شدند. به همین دلیل، جزئیات محاسبات این مسیر در ادامه ارائه نمی‌شود.

پیوست‌ها

اثبات لم ۱

اثبات. به دلیل $N_\epsilon \subset \mathcal{X}$ بدیهی است که $f^*_\epsilon \le f^*$. برای کران بالا، $x^* \in \arg\sup_{x\in\mathcal{X}} f(x)$ را در نظر بگیرید (یا دنباله‌ی بهینه‌ساز). طبق پوشش‌بودن، $\tilde{x}^* \in N_\epsilon$ وجود دارد به‌طوری که $d(x^*, \tilde{x}^*) \le \epsilon$. پس با لیپ‌شیتز بودن:

$$f(x^*) \le f(\tilde{x}^*) + L\epsilon \le f^*_\epsilon + L\epsilon.$$

چون $f^* = f(x^*)$، نتیجه حاصل می‌شود. □

اثبات لم ۲

اثبات. بازوهای بهینه را با $\mu^*_\epsilon := f^*_\epsilon$ و فاصله‌ها را با $\Delta_i := \mu^*_\epsilon - \mu_i \ge 0$ تعریف کنید. آنگاه

$$R^\epsilon_n = \sum_{i=1}^{K_\epsilon} \Delta_i\, \mathbb{E}[T_i(n)].$$

یک آستانه $\alpha > 0$ اختیار کنید و بازوها را به دو دسته تقسیم کنید.

دسته‌ی اول: $\Delta_i \le \alpha$. برای این بازوها داریم

$$\Delta_i T_i(n) \le \alpha T_i(n)$$

و لذا

$$\sum_{\Delta_i \le \alpha} \Delta_i \, \mathbb{E}[T_i(n)] \le \alpha \sum_{\Delta_i \le \alpha} \mathbb{E}[T_i(n)] \le \alpha n.$$

دسته‌ی دوم: $\Delta_i > \alpha$. برای این بازوها از $(\star)$ استفاده می‌کنیم:

$$\begin{aligned}\Delta_i \, \mathbb{E}[T_i(n)] &\le \Delta_i \Big(\frac{8\ln(n)}{\Delta_i^2} + 1 + \frac{\pi^2}{3} \Big) \\ &= \frac{8\ln(n))}{\Delta_i} + \Big(1 + \frac{\pi^2}{3}\Big)\Delta_i.\end{aligned}$$

چون $\Delta_i > \alpha$ و نیز $\Delta_i \le 1$ (زیرا $\mu_i \in [0,1]$)، نتیجه می‌شود:

$$\frac{8\ln(n)}{\Delta_i} \le \frac{8\ln(n)}{\alpha}, \qquad \Big(1+\frac{\pi^2}{3}\Big)\Delta_i \le \Big(1+\frac{\pi^2}{3}\Big).$$

پس

$$\sum_{\Delta_i > \alpha} \Delta_i \, \mathbb{E}[T_i(T)] \le \frac{8K_\epsilon \ln(n)}{\alpha} + \Big(1 + \frac{\pi^2}{3}\Big) K_\epsilon.$$

در مجموع:

$$R_n^\epsilon \le \alpha n + \frac{8K_\epsilon \ln(n)}{\alpha} + \Big(1 + \frac{\pi^2}{3}\Big) K_\epsilon.$$

اکنون $\alpha := \sqrt{\frac{8K_\epsilon \ln(n)}{n}}$ را انتخاب می‌کنیم تا دو جمله‌ی اول برابر شوند؛ آنگاه

$$\begin{aligned}\alpha n &= \sqrt{8K_\epsilon n \ln(n)}, \\ \frac{8K_\epsilon \ln(n)}{\alpha} &= \sqrt{8K_\epsilon n \ln(n)},\end{aligned}$$

و بنابراین

$$\begin{aligned}R_n^\epsilon &\le 2\sqrt{8K_\epsilon n \ln(n)} + \Big(1 + \frac{\pi^2}{3}\Big) K_\epsilon \\ &= 4\sqrt{2K_\epsilon n \ln(n)} + \Big(1 + \frac{\pi^2}{3}\Big) K_\epsilon.\end{aligned}$$

□

References

[۱] Peter Auer، Nicolò Cesa-Bianchi، and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*، ۲۵۶–۲۳۵:(۳–۲)۴۷ .۲۰۰۲،